NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. Government thereby incurs no responsibility, nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.

U. S. ARMY PERSONNEL RESEARCH OFFICE

An activity of the Chief, Research and Development

Charles S. Gersoni
Colonel, GS
Commanding

J. E. Uhlaner
Director, Research Laboratories

Hubert E. Brogden
Chief Scientist

Initial distribution of this report has been made by the U. S. Army Personnel Research Office. Additional copies may be obtained as follows. Department of Defense agencies and their contractors should address requests for additional copies to: Commander, Armed Services Technical Information Agency, ATTN: TIPDR, Arlington Hall Station, Arlington 12, Virginia. Other agencies and persons may obtain information on cost of microfilm or photostatic copies from: Office of Technical Services, Department of Commerce, Washington 25, D. C., ATTN: Technical Reports Section. Copies may also be obtained on loan from local depository libraries. A list of these libraries may be obtained from: Documents Expediting Project, Library of Congress, Washington 25, D. C. The mailing address of USAPRO is: U. S. Army Personnel Research Office, Washington 25, D. C.

AD 23/1, 28/4

U. S. Army Personnel Research Office, OCRD, DA
DEVELOPMENT OF ARMED FORCES QUALIFICATION TEST 7 AND 8 by
A. G. Bayroff and Alan A. Anderson. May 1963. Rept. on Input
Quality 00-01 Proj.--39 p. incl tables, figures 28 Ref.
(USAPRO Technical Research Report No. 1132)
(DA Project 2J024701A713) Unclassified Report

The Armed Forces Qualification Test, the screening test used by all the services, must provide both a measure of general military trainability and measures of specific aptitudes. Following the research design for previous forms, experimental test items in four content areas developed by the separate services were administered to 3000 Armed Forces personnel for item analysis and item selection. Final forms were then administered to standardization samples representative of the mobilization population as a basis for conversion of test scores to percentile norms. AFQT 7 and 8 correlated substantially with preceding operational forms (r = .89 - .90) and are satisfactory alternate forms for screening. Correlation of AFQT 7-8 with years of formal education (r = .53) was slightly less than for the previous forms. Because of the high degree of equivalence of the two forms (r = .94) established in samples totaling 600 cases, a single conversion table was established for AFQT 7 and 8. Based on experimentation, instructions for administering AFQT 7 and 8 have been made shorter and simpler than for previous forms, with no loss in test effectiveness.

UNCLASSIFIED
Human Resources Research
--Personnel Classification

AD 23/1, 28/4

U. S. Army Personnel Research Office, OCRD, DA
DEVELOPMENT OF ARMED FORCES QUALIFICATION TEST 7 AND 8 by
A. G. Bayroff and Alan A. Anderson. May 1963. Rept. on Input
Quality 00-01 Proj.--39 p. incl tables, figures 28 Ref.
(USAPRO Technical Research Report No. 1132)
(DA Project 2J024701A713) Unclassified Report

The Armed Forces Qualification Test, the screening test used by all the services, must provide both a measure of general military trainability and measures of specific aptitudes. Following the research design for previous forms, experimental test items in four content areas developed by the separate services were administered to 3000 Armed Forces personnel for item analysis and item selection. Final forms were then administered to standardization samples representative of the mobilization population as a basis for conversion of test scores to percentile norms. AFQT 7 and 8 correlated substantially with preceding operational forms (r = .89 - .90) and are satisfactory alternate forms for screening. Correlation of AFQT 7-8 with years of formal education (r = .53) was slightly less than for the previous forms. Because of the high degree of equivalence of the two forms (r = .94) established in samples totaling 600 cases, a single conversion table was established for AFQT 7 and 8. Based on experimentation, instructions for administering AFQT 7 and 8 have been made shorter and simpler than for previous forms, with no loss in test effectiveness.

UNCLASSIFIED
Human Resources Research
--Personnel Classification

AD 23/1, 28/4

U. S. Army Personnel Research Office, OCRD, DA
DEVELOPMENT OF ARMED FORCES QUALIFICATION TEST 7 AND 8 by
A. G. Bayroff and Alan A. Anderson. May 1963. Rept. on Input
Quality 00-01 Proj.--39 p. incl tables, figures 28 Ref.
(USAPRO Technical Research Report No. 1132)
(DA Project 2J024701A713) Unclassified Report

The Armed Forces Qualification Test, the screening test used by all the services, must provide both a measure of general military trainability and measures of specific aptitudes. Following the research design for previous forms, experimental test items in four content areas developed by the separate services were administered to 3000 Armed Forces personnel for item analysis and item selection. Final forms were then administered to standardization samples representative of the mobilization population as a basis for conversion of test scores to percentile norms. AFQT 7 and 8 correlated substantially with preceding operational forms (r = .89 - .90) and are satisfactory alternate forms for screening. Correlation of AFQT 7-8 with years of formal education (r = .53) was slightly less than for the previous forms. Because of the high degree of equivalence of the two forms (r = .94) established in samples totaling 600 cases, a single conversion table was established for AFQT 7 and 8. Based on experimentation, instructions for administering AFQT 7 and 8 have been made shorter and simpler than for previous forms, with no loss in test effectiveness.

UNCLASSIFIED
Human Resources Research
--Personnel Classification

AD 23/1, 28/4

U. S. Army Personnel Research Office, OCRD, DA
DEVELOPMENT OF ARMED FORCES QUALIFICATION TEST 7 AND 8 by
A. G. Bayroff and Alan A. Anderson. May 1963. Rept. on Input
Quality 00-01 Proj.--39 p. incl tables, figures 28 Ref.
(USAPRO Technical Research Report No. 1132)
(DA Project 2J024701A713) Unclassified Report

The Armed Forces Qualification Test, the screening test used by all the services, must provide both a measure of general military trainability and measures of specific aptitudes. Following the research design for previous forms, experimental test items in four content areas developed by the separate services were administered to 3000 Armed Forces personnel for item analysis and item selection. Final forms were then administered to standardization samples representative of the mobilization population as a basis for conversion of test scores to percentile norms. AFQT 7 and 8 correlated substantially with preceding operational forms (r = .89 - .90) and are satisfactory alternate forms for screening. Correlation of AFQT 7-8 with years of formal education (r = .53) was slightly less than for the previous forms. Because of the high degree of equivalence of the two forms (r = .94) established in samples totaling 600 cases, a single conversion table was established for AFQT 7 and 8. Based on experimentation, instructions for administering AFQT 7 and 8 have been made shorter and simpler than for previous forms, with no loss in test effectiveness.

UNCLASSIFIED
Human Resources Research
--Personnel Classification

Technical Research Report 1132

DEVELOPMENT OF ARMED FORCES QUALIFICATION TEST 7 AND 8

A. G. Bayroff and Alan A. Anderson

Submitted by Edmund F. Fuchs
Chief, Military Selection Research Laboratory

Approved by

JEUhlaner
J. E. Uhlaner
Director, Research Laboratories

H E Brogden
H. E. Brogden
Chief Scientist

Charles S. Gersoni
Charles S. Gersoni
Colonel, GS
Commanding

Army Project Number
2J024701A713

Input Quality 00-01

May 1963

USAPRO Technical Research Reports and Technical Research Notes are intended for sponsors of R&D tasks and other research and military agencies. Any findings ready for implementation at the time of publication are presented in the latter part of the Brief. Upon completion of a major phase of the task, formal recommendations for official action normally are conveyed to appropriate military agencies by briefing or Disposition Form.

PREFACE

The present publication reports on a part of the INPUT QUALITY Task, which is responsive to special requirements of the Department of Defense AFES Policy Board, the Assistant Secretary of Defense (Manpower), and the Deputy Chief of Staff for Personnel.

Congressional legislation has provided the basis for procedures to screen input quality so that those who lack military trainability may be rejected. Successive forms of the Armed Forces Qualification Test (AFQT) meet the requirements in the form of an overall screening measure. The concept of a common instrument to be used in screening inductees and enlistees for all the services was originated by Dr. J. E. Uhlaner in the late 1940's. The concept was based on the recognized principle that requirements for initial enlistment and induction into the several services are common with respect to basic mental abilities. These requirements were conceived of first as verbal ability, arithmetic reasoning ability, and spatial relations ability basic to mechanical and automotive jobs, and later as including understanding of tool functions. The common screen for all the services has the advantage of providing an objective basis for the equitable distribution of manpower with respect to general military trainability.

Development of successive AFQT forms has been in accordance with the direction of the Armed Forces Examining Station Policy Board, Department of Defense, and in close coordination with representatives of policy and operating agencies of the Services. In these developments, the Army in its capacity as Executive Agent of the Board has had overall responsibility, with the Navy, the Air Force, and the Marine Corps contributing substantial effort. USAPRO Technical Research Report 1132 describes research involved in developing the most recent forms of the AFQT.

DEVELOPMENT OF ARMED FORCES QUALIFICATION TEST 7 AND 8

BRIEF

Requirement:

Successive forms of the Armed Forces Qualification Test (AFQT) are developed to meet the requirement set by Congress for procedures to screen selective service registrants for military trainability and to meet the needs of the armed services for means of screening potential enlistees.

~~Procedure~~:

Experimental test items developed by the separate services were administered experimentally to 3000 Armed Forces personnel. Final forms developed on the basis of experimental data were then administered to standardization samples representative of the mobilization population as a basis for conversion of test scores to percentile norms.

~~Findings:~~

AFQT 7 and 8 are in substantial agreement with preceding operational forms and are satisfactory alternate forms with respect to screening function. The two forms are highly equivalent. A single table for converting raw scores to percentile norms applies to both.

AFQT 7 and 8 scores are no more dependent upon amount of formal education than were scores on previous forms.

Based on experimentation, instructions for administering AFQT 7 and 8 have been made shorter and simpler than for previous forms with no loss in test effectiveness.

Utilization of Findings:

Operational AFQT 7 and 8 provide both a measure of general trainability and measures of specific aptitudes corresponding to certain of the Army aptitude area scores.

CONTENTS

TABLES Page

FIGURES

DEVELOPMENT OF ARMED FORCES QUALIFICATION TEST 7 AND 8

BACKGROUND

The Armed Forces Qualification Test (AFQT) is the primary screening test used by all the services to determine mental acceptability of applicants for enlistment and selective service registrants. New forms--AFQT 7 and 8--were introduced 1 July 1960. Introduction of the new forms completed three years of work by research personnel of the Army, Navy, Air Force, and Marine Corps, the Army having major responsibility for planning and executing the research as well as for implementing the products.

The new forms are the latest in a series of tests initiated 1 January 1950 for the purpose of providing uniform mental ability screening for all services. The series was initiated as a result of Congressional legislation in 1948 (PL 759, 80th Congress) and in 1951 (PL 51, 82d Congress) and directives of the Secretary of Defense.

Successive AFQT Forms

Forms 1 and 2 of the AFQT replaced the variety of screening tests and scoring systems which had been used by the several services. A single test had one great advantage--it provided a basis for a more equitable distribution of mental quality among the services. Since 1950, several new forms have been introduced--AFQT 3 and 4 on 1 January 1953, and AFQT 5 and 6 on 1 August 1956. Work on AFQT 7 and 8 was formally initiated by a memorandum from the Chairman, Armed Forces Examining Station Policy Board to the Army Member, subject: Development of Forms 7 and 8 of the AFQT, dated 5 March 1957.

Operational tests need to be revised or replaced when they have been seriously compromised, when content has become outdated, or when research findings point the way to improvement in the test. New alternate forms of the AFQT--most widely used of all military tests--are prepared routinely at set intervals because of the possibility of any or all of these contingencies. Outdated content of aptitude tests--as distinguished from proficiency tests--does not necessarily result in reduced effectiveness of the test. However, outdated content may arouse unfavorable attitudes toward the test, and retention of such content is therefore undesirable. Minor editorial revisions may be made without reinvestigating the research basis. In the case of new forms entailing a complete change of content, as with AFQT 7 and 8, the complete research cycle is required.

The Research Cycle

All forms of AFQT, including AFQT 7 and 8, have been the product of a research cycle consisting of the following phases prior to implementation:

1. Construction of experimental tests
2. Administration of experimental tests
3. Statistical analysis of experimental test data
4. Selection of test items for the new forms
5. Standardization of the new forms

At two phases--tryout of the experimental tests and standardization of the final forms--large numbers of personnel must be tested experimentally at Armed Forces Examining Stations, reception centers, and training centers. Installations of all the services furnish examinees for these purposes. Implementation of the new tests is not, properly speaking, a research phase. However, research personnel, because of their intimate knowledge of both technical and operating requirements, have an indirect but major responsibility for putting the new forms into use.

THE EXPERIMENTAL TESTS

Planning The Tests

A conference of research personnel of the Army, Navy, Air Force, and Marine Corps was held 13 June 1957 to plan development of the new tests.[1] The conference agreed that the new forms--like all previous forms--should

1. provide a measure of overall ability and permit deriving measures of more specific aptitudes if desired

2. not be excessively dependent on amount of formal education

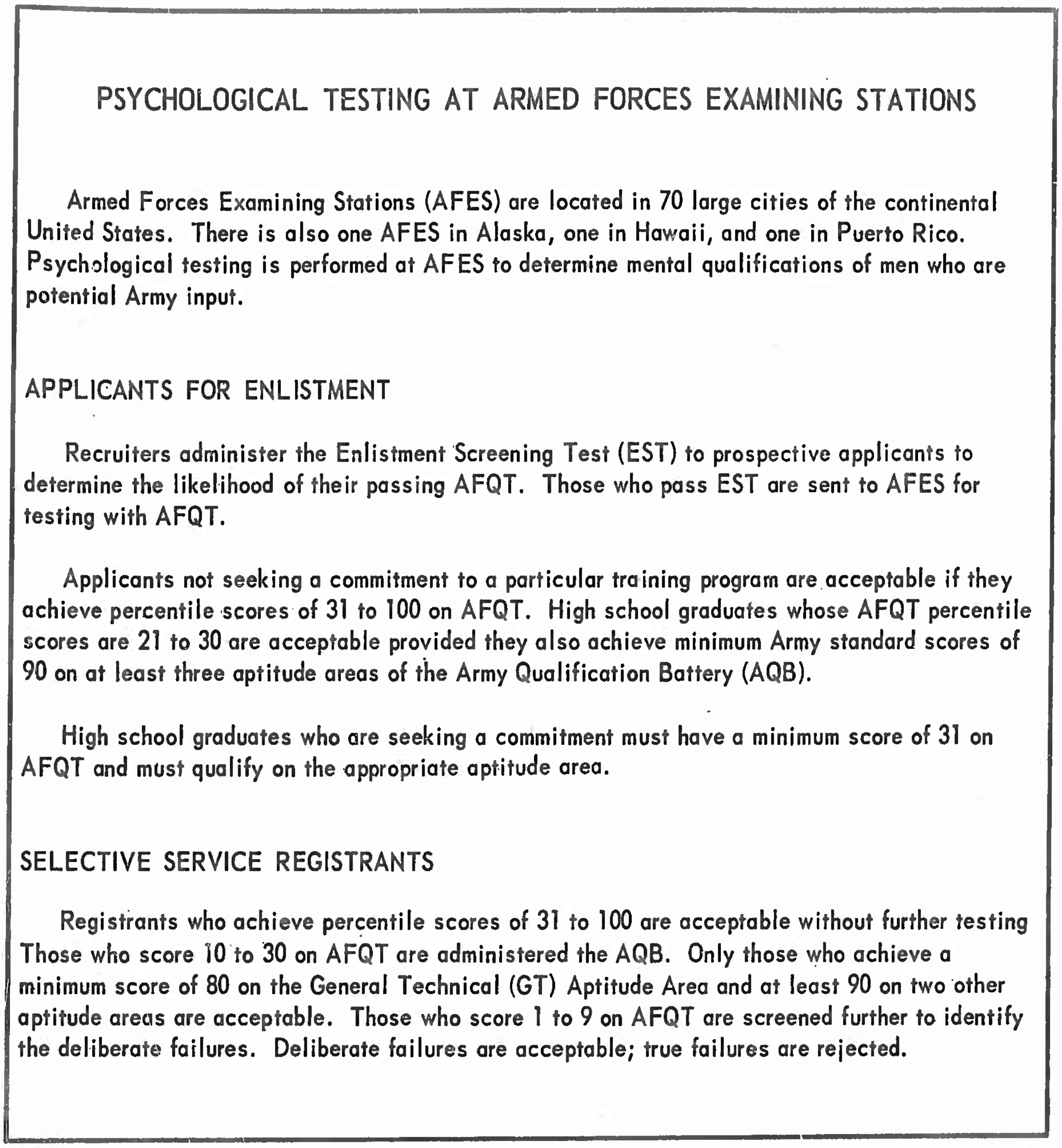
3. be of appropriate difficulty for a wide-range population and at the same time have special sensitivity in critical score zones

4. emphasize power rather than speed

5. permit uniform test administration within one hour to large groups of examinees with widely differing backgrounds.

The new forms, as prior forms, were to contain equal proportions of items measuring verbal ability, arithmetic reasoning, understanding of tool functions, and perception of spatial relations (Figure 2). Each form was to consist of 100 items, arranged in order of difficulty. Responsibility for construction of items was allocated among the services.

[1] Substantial contributions were made to specific phases by the following research personnel: ARMY--Mary A. Morton, Richard H. Hilligoss; NAVY--Charles I. Hodges; AIR FORCE--Jane McReynolds; MARINE CORPS--Francis F. Medland.



PSYCHOLOGICAL TESTING AT ARMED FORCES EXAMINING STATIONS

Armed Forces Examining Stations (AFES) are located in 70 large cities of the continental United States. There is also one AFES in Alaska, one in Hawaii, and one in Puerto Rico. Psychological testing is performed at AFES to determine mental qualifications of men who are potential Army input.

APPLICANTS FOR ENLISTMENT

Recruiters administer the Enlistment Screening Test (EST) to prospective applicants to determine the likelihood of their passing AFQT. Those who pass EST are sent to AFES for testing with AFQT.

Applicants not seeking a commitment to a particular training program are acceptable if they achieve percentile scores of 31 to 100 on AFQT. High school graduates whose AFQT percentile scores are 21 to 30 are acceptable provided they also achieve minimum Army standard scores of 90 on at least three aptitude areas of the Army Qualification Battery (AQB).

High school graduates who are seeking a commitment must have a minimum score of 31 on AFQT and must qualify on the appropriate aptitude area.

SELECTIVE SERVICE REGISTRANTS

Registrants who achieve percentile scores of 31 to 100 are acceptable without further testing Those who score 10 to 30 on AFQT are administered the AQB. Only those who achieve a minimum score of 80 on the General Technical (GT) Aptitude Area and at least 90 on two other aptitude areas are acceptable. Those who score 1 to 9 on AFQT are screened further to identify the deliberate failures. Deliberate failures are acceptable; true failures are rejected.

Figure 1. Screening for Enlisted Input Quality

It was a small table.

(A) STURDY

(B) ROUND

(C) CHEAP

(D) LITTLE

A boy buys a sandwich for 20 cents, milk for 10 cents, and pie for 15 cents. How much did he pay in all?

(A) 30 CENTS

(B) 35 CENTS

(C) 45 CENTS

(D) 50 CENTS

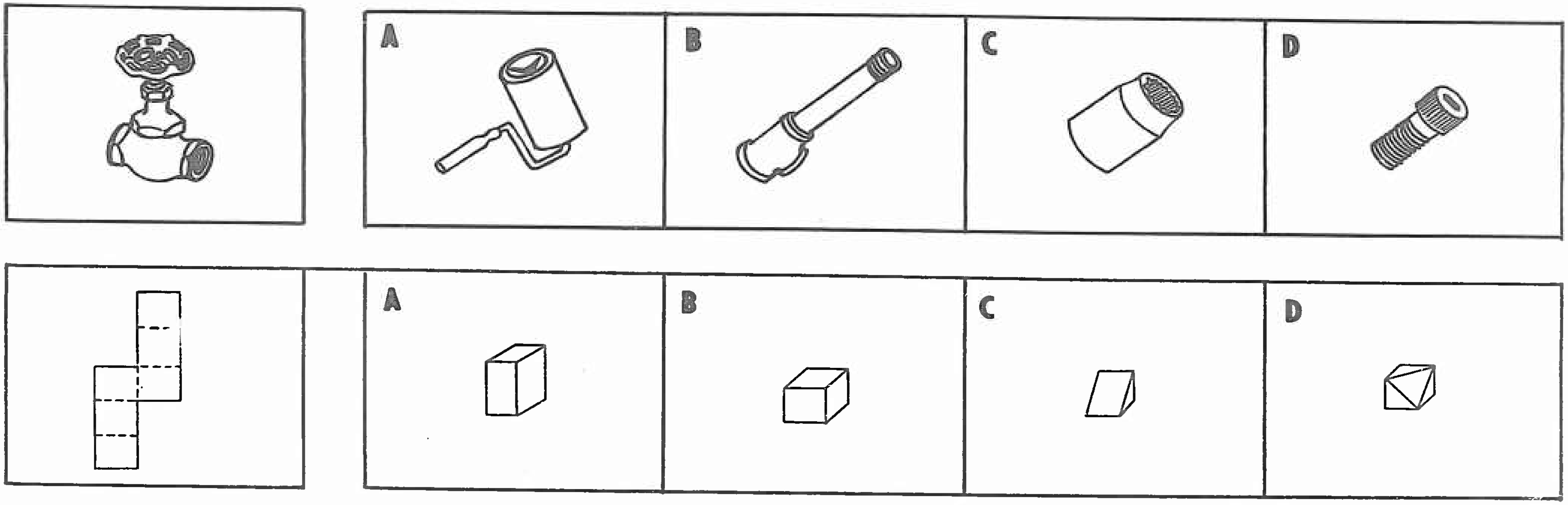


Figure 2. Sample Items--Experimental AFQT Forms.

Construction of the Experimental Tests

The experimental tests were essentially item pools from which items were to be selected for use in the operational forms. Experimentation is necessary because test items frequently fail to live up to expectations despite the best efforts of experienced item constructors. Also, it is necessary to determine how difficult the items are in order to match items for the two parallel forms. Hence, the experimental test contained a much larger number of items than were to be used. Two tests of 300 items each were prepared for experimental administration, with 75 items in each of the four content areas. Items from each content area were arranged in successive blocks throughout the tests so that even the slower examinees would have a chance to respond to some items in each area.

ADMINISTRATION OF EXPERIMENTAL TESTS

The experimental tests were administered to 3,000 examinees at eight training centers, among which all services were represented, and at seven Armed Forces Examining Stations. Locations were chosen to provide a wide geographic sampling so that item statistics would be broadly based and not reflect characteristics of a particular region. Research personnel of all the services were present at the selected installations to administer the tests and establish the samples.

Installations at which experimental tests were administered to provide data on which to select items for final forms of AFQT 7 and 8.

SERVICE	TRAINING CENTERS	AFES
Army	Fort Jackson, S. C. Fort Knox, Ky. Fort Carson, Colo.	Columbia, S. C. Louisville, Ky. Denver, Colo.
Navy	Great Lakes, ILL. San Diego, Calif.	Chicago, ILL. Oakland, Calif.
Air Force	Lackland AFB, Tex.	San Antonio, Tex.
Marine Corps	Parris Island, S. C. San Diego, Calif.	Richmond, Va.

To provide data that would reflect the more stable mobilization population rather than a particular input sample, examinees were selected on the basis of their current AFQT 5 and 6 scores in proportions corresponding to the distribution of scores on comparable tests for samples obtained during the peak of mobilization in World War II. At that time the strength of the services was 12,000,000 enlisted men and commissioned officers.

In research for previous AFQT forms, examinees for the mobilization samples had been selected by actually administering a test similar to the one used in World War II. In the case of AFQT 7 and 8, examination of the technical problems involved led to a decision to omit the extra test and use the already available scores on AFQT 5 and 6. This procedure could result not only in a technical advantage but also in a considerable saving of time and money.

SELECTION OF ITEMS FOR THE FINAL FORMS

The next phase was the statistical analysis of the test data. This analysis provided information on the effectiveness of each of the 600 experimental test items from which the final forms were to be assembled. The statistical analysis required a considerable amount of time, and could not have been completed within a reasonable time limit without the aid of conventional statistical machines and also of a special purpose electronic computer.

The item analysis yielded a variety of information: The difficulty of the item was indicated by the number of examinees who responded correctly. Another important type of information was the relation between response to an item and response to other items in the same content area ("internal consistency"). These two types of item statistic, along with other information, were used in evaluating the items so that appropriate selection could be made from the pool of experimental items.

Selection of items was guided by a number of considerations. First, the level of difficulty had to be appropriate for a wide-range population, and at the same time afford special sensitivity in critical score zones. Accordingly, items were selected so as to provide the appropriate number for each level of ability--the minimum number at the extremely high and extremely low levels, and a greater number at the levels where most of the screening would occur, that is, in the zone represented by the 10th - 31st percentiles. A second important consideration was the pattern of internal consistency indexes; to be selected, an item not only had to be substantially correlated with other items of the same content area, but this correlation should be higher than correlation with items of the other content areas. Finally, the item analysis provided a partial check on ambiguity of items, particularly those that were not extremely difficult or extremely easy: The correct answer should be selected more often than any other answer in the same item. In addition, the items selected had to meet the general requirement of appropriate simplicity and clarity. Further details of item selection are presented in the Technical Supplement.

Items were selected in matched pairs, one for each of the two final forms, to insure that the two forms would be highly comparable. The forms prepared for standardization each consisted of 100 items, 25 in each content area. Items were arranged in order of difficulty, from easiest to hardest, to provide a power test. Each level of difficulty contained items of all four content areas. Standard instructions to examinees and to examiners, based on experience with instructions for the experimental forms, were prepared. The final forms were then used in a dual study: (1) the standardization proper, that is, the conversion of the raw scores to percentile scores; and (2) the measurement of the degree of equivalance of the two forms.

STANDARDIZATION OF FINAL FORMS

Purposes of Standardization

Regardless of how well items have been selected for a test, the process is not complete. It is still necessary to establish what a score on the test (the "raw" score) means. For AFQT, this means calibrating the relative standing in the mobilization population indicated by a given raw score. Raw scores are converted into mobilization percentile scores. Under acceptance standards set by the Congress, it is necessary to be able to say that a raw score of x points indicates that the examinee has done as well as the lowest 10 percent of the mobilization population. The percentile score of 10 must have the same meaning in any form of AFQT, even if the forms differ in difficulty, length, or content. For example, a raw score of 24 on AFQT 5 and 6 and a raw score of 27 on AFQT 1 and 2 are both equivalent to a percentile score of 10.

To accomplish this conversion, the sample on which the new forms of the tests are standardized must be representative of the mobilization population, so that when the standardized tests are administered operationally the percentage of examinees attaining particular scores will be the same as in the mobilization population. This equivalence holds only if applied to scores made by individuals who are being evaluated as members of the population on which the test was standardized. This is an important requirement, and unless it is met, the number of failures in any preinduction input will not amount to the 10 percent permissible under the passing score set by Congress.

An example will clarify this point. During peacetime, deferment policies are liberal. In effect, substantial numbers of eligible men are not ordered up for preinduction processing. By far most of the men deferred are of ability levels well above the passing score on AFQT (college students who maintain the necessary academic grades, ROTC students, men in critical occupation, etc.). As a result, failures constitute a greater proportion of the men who are called up than they would in a typical mobilization population. In addition, many enlistees come out of the selective service pool. Since these enlistees have had to attain AFQT scores above the 10th percentile, the proportion above the 10th percentile in a given preinduction input is likely to be further reduced. Further, the actual percentage of rejections for Army service has been increased by the requirement that, to be acceptable, preinduction examinees who obtain AFQT percentile scores 10 - 30 must be screened further (as of 1 May 1963, they must achieve an Army standard score of 80 on the General Technical (GT) Aptitude Area and a minimum of 90 on at least two other aptitude areas).

The AFQT differs from conventional intelligence tests both in its more restricted purpose of measuring military trainability, and in the fact that normative data for all forms have been obtained on samples representative of one segment of the nation's population--males eligible for military service at the peak of mobilization during World War II. In other words, the test shows how a man compares with other potential servicemen in the ability to learn and perform in military jobs to which he may be assigned rather than how he compares with other individuals of his own age in general mental capacity.

Administration of Standardization Forms

The final forms of the test were administered to examinees at nine training installations of the Army, Navy, Air Force, and Marine Corps and at eleven Armed Forces Examining Stations selected so as to provide wide geographic sampling. Again, research personnel of all the services participated in the field work. As a basis for establishing converted scores, each form was administered to 1800 examinees. To ascertain the equivalence of the two alternate forms, both forms were administered to an additional 1000 examinees.

Installations at which standardization forms and equivalence studies of AFQT 7 and 8 were conducted.

SERVICE	TRAINING CENTERS	AFES
Army	Fort Dix, N. J. Fort Jackson, S. C. Fort Knox, Ky. Fort Carson, Colo.	New York, N. Y. Columbia, S. C. Louisville, Ky. Denver, Colo.
Navy	Great Lakes, ILL. San Diego, Calif.	Chicago, ILL. Oakland, Calif.
Air Force	Lackland AFB, Tex.	Dallas, Tex. Houston, Tex. San Antonio, Tex.
Marine Corps	Parris Island, S. C. San Diego, Calif.	Baltimore, Md. Atlanta, Ga.

Mobilization samples were established for both studies in the same way as for the experimental tests. Cases were selected according to their operational AFQT 5-6 scores in such proportions as to reproduce the distribution of comparable scores during peak mobilization of World War II. For the standardization procedure, an additional test was administered. This test, an editorial revision of the Army General Classification Test (AGCT) used to define the World War II mobilization population, provided the percentile norms to which raw scores of the AFQT were to be converted. In the development of previous forms of the AFQT, this revised World War II test (R 9) had been used to select cases for the sample as well as to provide the percentile norms--a time-consuming procedure, since answer sheets had to be scored before cases could be selected. Substitution of the operational AFQT as a basis for selecting cases not only decreased field and research costs, but more important, also permitted an increase in stability of the norms, derived, as before, with the use of R 9. The procedure was possible in part because operational administration and scoring of the AFQT hes in recent years been well controlled.

Statistical Evaluation of the New Forms

Statistical analysis of data obtained in the standardization samples showed that both AFQT 7 and 8 agreed substantially with the current AFQT 5 and AFQT 6 (the test used to constitute the samples) and with R 9 (the test used to establish the mobilization percentile equivalents). The equivalence data showed AFQT 7 and AFQT 8 to be highly comparable.

Performance on AFQT 7 and 8 was found to be less dependent on number of years of schooling than was the case with the earlier forms. Performance on such tests as AFQT cannot be completely independent of amount of formal schooling. Additional education affects an individual's abilities; also, the more capable individuals are encouraged to continue their schooling longer. However, other factors--such as legal age limits for compulsory schooling, community customs, automatic promotion to keep a child with his own age group ("social promotion")--tend to reduce the significance of years of education. It is therefore not safe to infer from the number of years of schooling a person has had either the level of his ability or the amount he has learned.

Conversion of Raw Scores to Mobilization Percentile Scores

Tables for converting raw scores to mobilization percentile scores were developed separately for AFQT 7 and AFQT 8. For each raw score (number right corrected for chance success), the percentile assigned was the R 9 percentile score which has the same cumulative frequency. In this manner, tables were constructed showing the percentile in a mobilization population equivalent to each raw score on AFQT 7 and AFQT 8. After necessary statistical adjustments, tables for the two forms were compared. The tables proved to be substantially alike, permitting use of one conversion table for both forms of AFQT--the same raw score on either form represents the same percentile score. A single table was possible only because the two forms were highly comparable and the conversions very similar.

RELATED RESEARCH

Following the development of the tests, materials were prepared for printing and distribution. In addition, research was completed on auxiliary devices which were to be implemented concurrently with AFQT 7 and 8. This research, which is reported separately, led to the development of scoring keys for determining literacy level of AFQT failures and for identifying deliberate failures on AFQT. On 1 July 1960, the new forms were introduced for operational use.

Paralleling this developmental research, other research conducted by USAPRO has a bearing on AFQT forms. Two research studies in particular have affected the development of AFQT 7 and 8. One involved revision of the reference test used in the standardization phase to provide the basis for equating scores on the new forms of AFQT to scores on the earlier forms and to permit interpreting the scores in terms of the mobilization population. The second research study indicated that the instructions for taking AFQT could be considerably simplified without introducing serious error in the test scores.

Research is continuing, which, if successful, will have an impact on future forms of AFQT and related instruments. Among the problems under long-range investigation are the following: (1) possibilities of reducing the effective length of tests without impairing their screening effectiveness; (2) usefulness of new content in further increasing screening effectiveness; (3) possibilities of more economical ways of developing methods to identify deliberate failures; (4) determination of the currency of the test definition of the mobilization population; (5) promise of programmed testing (including use of machines) for increasing screening effectiveness through the use of techniques now impossible.

Selected USAPRO Publications
Bearing on INPUT QUALITY Research

Bayroff, A. G. Methods for improving enlisted input--Status report, 30 June 1962. Technical Research Report 1125. June 1962.

Bayroff, A. G. The mobilization base for AFQT Norms. Research Memorandum 63.8. May 1963.

Bayroff, A. G., and Anderson, A. A. Development of literacy screening scales for AFQT 7 and 8 failures. Technical Research Note 131. January 1963.

Bayroff, A. G. Successive AFQT Forms--Comparisons and evaluations. Technical Research Note 132. May 1963.

Bayroff, A. G., Heermann, E. F., and Anderson, A. A. Screening devices for selective service registrants who fail AFQT 7 and 8. Technical Research Report 1130. January 1963.

Bayroff, A. G., Fuchs, E. F., and Seeley, L. C. New instruments for screening Army input. American Psychological Association meeting, 7 September 1960. (Abstr.) *The American Psychologist*, vol 15, pp. 496-97. 1960.

Bayroff, A. G., Mundy, J. P., and Uhlaner, J. E. Development of new forms of the Army Qualification Test, AFQT 5 and AFQT 6. American Psychological Association meeting, 4 September 1956. (Abstr.) *The American Psychologist*, vol 11, p. 427. 1956.

Bayroff, A. G., Seeley, L. C., and Anderson, A. A. Development of the Army Qualification Battery, AQB-1. Technical Research Report 1117. October 1959.

Bayroff, A. G., Seeley, L. C., and Anderson, A. A. Relationship of AFQT to rated basic training performance. Technical Research Note 106. February, 1960.

Bayroff, A. G., Thomas, J. A., and Kehr, Carol J. Evaluation of EST for predicting AFQT performance. Technical Research Report 1114. February 1959.

Bolanovich, D. J., Harper, Bertha, Birnbaum, A. H., Lovelace, N. R., and Uhlaner, J. E. Development and standardization of short forms of the Armed Forces Qualification Test. Technical Research Report 934. April 1952.

Bolanovich, D. J., Jones, W., Lovelace, N. R., and Uhlaner, J. E. Follow-up of the standardization of the Armed Forces Qualification Test. Technical Research Report 956. June 1952.

Bolanovich, D. J., Mundy, J. P., Jones, W., Klieger, W. A., Houston, T. J., and Marks, M. R. Test performance of administrative inductees. Technical Research Report 1080. October 1953.

Morton, Mary A., Houston, T. J., Mundy, J. P., and Bayroff, A. G. Mental screening tests for women in the armed forces. Technical Research Report 1103. May 1957.

Schenkel, K. F., Burke, Laverne, K., and Marks, M. R. Construction of the Examen Calificación de Fuerzas Armadas. Technical Research Report 1090. December 1954.

Schenkel, K. F., and Fuchs, E. F. Development of an aptitude test for a foreign population. American Psychological Association meeting. 4 September 1956. (Abstr.) The American Psychologist, vol 11, p. 427. 1956.

Schenkel, K. F., Meyer, L. A., Rosenberg, N., and Bayroff, A. G. Evaluation of the Puerto Rican screening test (ECFA) against success on the job. Technical Research Report 1106. June 1957.

Seeley, L. C., Morton, Mary A., and Anderson, A. A. Exploratory study of a sequential item test. Technical Research Note 129. December 1962.

Uhlaner, J. E. Development of the Armed Forces Qualification Test and predecessor Army screening tests, 1946-1950 (3d printing). Technical Research Report 976. November 1952.

TECHNICAL SUPPLEMENT

TECHNICAL SUPPLEMENT

ITEM ANALYSIS AND ITEM SELECTION

The Experimental Tests

The experimental tests AFQT 7-8X and AFQT 7-8Y[2] were essentially item pools from which items were to be selected for operational use. These pools were intended to allow for the usual attrition after item analysis and to provide flexibility in matching items for two parallel forms, as well as to develop a residue of tested items for possible use in new forms of the Enlistment Screening Test (EST). The EST is administered by recruiters to applicants for enlistment in order to reject those who are likely to be rejected by the longer AFQT.[3] Experience has indicated that a pool of 600 items is generally adequate to provide at least 100 pairs of matched items with the required characteristics. A possible exception was the tool functions area--attrition of tool functions items has been greater than attrition of items in the other content areas. Each experimental test consisted of 75 items in each of the four content areas.

In the past, three experimental tests each consisting of 200 items have been prepared for administration to three samples. However, reductions in enlisted strength made it desirable that the current test development be conducted using a smaller number of examinees than previously. It was decided to use a reduced number of examinees and to constitute two comparable samples, one for each of two experimental tests. This procedure resulted in two 300-item tests. It was expected that the greater length would not seriously affect the reliability of the item statistics inasmuch as the latter part of the test contained the more difficult items, items which the less able examinees would not be expected to answer correctly except by chance. Since a brief rest period was introduced and additional time was allowed, the performance of the abler examinees was not expected to suffer.

[2] The following practice has been adopted throughout the present report to designate treatment of alternate forms: where no distinction is made between forms, the hyphen is used as in "AFQT 5-6"; otherwise, the two forms are listed separately, "AFQT 5 and 6", for example.

[3] For information on the current forms of the Enlistment Screening Test, see Morton, Houston, and Bayroff, 1957; and Bayroff, Thomas, and Kehr, 1959.

Sampling

Sampling design. The design for administration of the experimental tests was consistent with the ultimate requirement of developing national norms. Caser were selected to represent the range of relevant ability to be found in the mobilization population. Geographical sampling was employed so that item statistics would be broadly based and not reflect the characteristics of a particular region. Armed Forces Examining Stations (AFES) and training installations of the Army, Navy, Air Force, and Marine Corps located in five of the six Army areas were involved in the testing (page 5). The design was consistent also with the ultimate requirement that norms be based on the distribution of ability in the mobilization population (Bayroff, 1963). Mobilization samples for testing were established by selecting cases so that the distribution of percentile scores on AFQT 5-6 reproduced the distribution of percentile score equivalents on the Army General Classification Test (AGCT) during the peak of mobilization in World War II. The procedure was based on the fact that AFQT percentile scores are derived from the mobilization distribution of Army standard scores on the reference test AGCT. An equal number of cases in each AFQT decile was tested with each of the two experimental forms.

The use of the operational AFQT as the sampling test made it unnecessary to administer and score an additional test before the samples could be constituted. Several considerations made this economy possible: (1) In recent years, operational administration of AFQT has markedly improved; (2) AFQT 5 and 6 had been standardized with Classification Test R 5 (an early form of AGCT) as the reference test; and (3) The correlation between operational AFQT 5 and 6 scores and R 5 scores was substantial (r = .84, .86).

Data collection procedures. The two experimental tests were administered during June 1958 to a total of 3015 examinees representing all the services so as to insure the availability of 1000 complete cases, appropriately distributed, for each form. Testing sessions were 3-1/2 hours long, with three hours allotted for the test proper, and a ten-minute rest period after the first hour. The tests were administered by the regularly assigned examining personnel under the supervision of research personnel. Standard testing conditions were observed. Quotas of examinees for each AFQT decile were established for each location.

Training centers were the source of the bulk of the cases. The experimental tests were administered to enlistees and inductees who had scored above the 9th percentile on AFQT (that is, percentiles 10 - 100). Since only the Army was accepting inductees (Selective Service registrants who had attained AFQT percentile scores of 10 or higher), all the examinees in the percentile interval 10 - 19 were supplied by the Army. Examinees were tested no later than their second week of basic combat training, one to two weeks after classification testing, and a varying interval--which might be as long as several months--after operational AFQT testing at the AFES.

At AFES, the experimental tests were administered only to those Selective Service registrants undergoing preinduction processing who were identified as true failures on the operational AFQT (according to the Terminal Screening Guide, Department of the Army Pamphlet 611-37, July 1956). By this procedure cases in percentiles 1 - 9 were obtained. No applicants for enlistment were tested at AFES since the Enlistment Screening Test had previously been administered to reject those who would not meet the AFQT qualifying score. The qualifying score for enlistment varied among the services, but in every case was higher than the 10th percentile.

Selection of cases for analysis. From the 1500 cases collected for each of the two experimental forms, 1000 cases were selected for the item analysis. To reproduce the mobilization population distribution, proportional numbers of cases were drawn from each location and from each service, after which 100 cases were selected for each operational AFQT decile in such a manner that each decile contained as rectangular a distribution as possible. The number of cases available was sufficiently large to require only occasional compromises.

Each case selected was expected to be complete and to contain a minimum number of omitted items (arbitrarily set at no more than 15). However, examination revealed that it would be necessary to accept a number of cases with more than the intended minimum of omitted items, if the required frequencies in the two lowest AFQT deciles[4] were to be attained. Only cases in which the omissions were terminal (unmarked items after the last marked item) were considered for addition to the sample. It was then necessary to make adjustments to permit the computation of item statistics for the sample including terminal omission cases.

Adjustment for omitted items. The adjustments (Bayroff, Morton, Anderson, and Hilligoss, 1960) made in the two lowest deciles were based on the following assumptions: (1) Since the items in each experimental test were arranged in order of increasing estimated difficulty, the terminal items could not have been answered correctly by the low level examinees except by chance; hence, had these examinees completed the test, the average increase in score, after correction for chance success, would have been zero. (2) Had these examinees completed the test, their pattern of responses would have been the same as the pattern of examinees of comparable AFQT level who did complete the test. (3) The examinees who did not complete the test received the same mean corrected score (R - W/3) on each content area as did comparable examinees who completed the test.

[4] Strictly speaking, the lowest interval was not a full decile. Its range was percentile scores 1 - 9, the fa·lures, and it contained 90 cases. The next interval, percentile scores 10 - 20, contained 110 cases. A percentile score of 0 is administratively assigned to indicate no examination.

To verify the assumptions made, the experimental test performance of low-level examinees who did not complete the test was compared with the performance of comparable examinees who completed the test (Table 1). No differences between the two low-level groups were found in total score or in three of the four content area scores (Anderson, 1960). The only significant difference (CR = 2.74) found was in the spatial relations area, where the examinees who did not complete the test obtained higher mean scores than did those who completed the test. Thus, the assumptions made appear not to have been as appropriate for the spatial relations area as for the other content areas, illustrating the possible bias that may be introduced in item analysis samples if cases of terminal ommissions are excluded from the sample.

Table 1

COMPARISON OF SCORES OF TERMINAL OMISSION EXAMINEES WITH SCORES OF EXAMINEES WHO COMPLETED THE TEST

	Verbal		Arithmetic Reasoning		Tool Functions		Spatial Relations[a]		Total Score	
	M	SD	M	SD	M	SD	M	SD	M	SD
Terminal Omissions	16.2	12.4	15.6	9.6	22.8	11.6	19.5	12.7	72.2	33.1
Completed Test	18.9	14.6	13.4	12.4	24.4	12.3	14.7	12.2	71.4	35.2

[a] Critical ratio of the difference between means was 2.74.

Item Statistics Computed

Item statistics were computed separately for each of the two experimental forms and were based on the adjusted values occasioned by the inclusion of terminal omission cases in the two lowest deciles.

Difficulty index. For each item, raw p-value (frequency of correct response) was corrected for chance success according to the formula

$$\frac{R - \frac{W}{3}}{R + W}, \text{ with } R + W = 1000.$$

The corrected p-value was expressed as a decimal.

Internal consistency. The biserial coefficient of correlation was computed between pass-fail on each item and the total score on all 75 items in the same content area in a given experimental test.

Independence. Biserial correlation coefficients were computed between pass-fail on each item in one content area and total score on each of the other three content areas. Degree of independence of a content area was estimated by obtaining the mean of these biserial coefficients.

Relationship with operational AFQT score. Biserial correlation coefficients were computed between pass-fail on each item and total score on operational AFQT. Means of these coefficients were computed for each content area.

Selection of Items for Standardization Forms

The statistics computed for all items were used as guides in the selection of the 200 most effective items from the pool of 600 experimental items. The following criteria guided the item selection:

1. The p-value for the correct response should be greater than for any incorrect response in the same item.

2. Incorrect responses should have equal p-values. No statistical tests of the significance of differences in p-values were applied.

3. Biserial coefficients of correlation should be high with total score on all items of the same content area and lower with total scores on the other content areas. Correlation with other content areas could not be too low since this would introduce the possibility of selecting items of inadequate reliability.

4. Correlation with total operational AFQT score should be substantial.

The above specifications could not be adhered to with complete objectivity in all cases. Furthermore, use of biserial correlation coefficients had to take difficulty level into account, since the very easy or very difficult items showed extreme p - q splits in the distribution which could produce highly unstable coefficients.

One other statistical specification had to be met, namely, the distribution of item difficulty indexes. This distribution was established through consideration of the dual use of AFQT: (1) its use as the basis for equitable allocation of ability among the services, which required a wide range of item difficulty, and (2) its use as a screening instrument, which required a greater proportion of items whose difficulty was appropriate to the range (10th - 31st percentiles) where maximal discrimination was required. There was no need to discriminate within the highest quartile or below the 5th percentile.

The distribution of item difficulty indexes (corrected p-values) arrived at and the corresponding AFQT percentile intervals are shown in Table 2. The number of items desired for each AFQT interval within each content area of 25 items was established according to the discriminations required. Each entry in column 2 shows the percentage of the mobilization population that can be expected to know the correct response, that is, the raw p-value corrected for chance success. Items were selected so that the distribution of corrected p-values approached the desired distribution as closely as possible.

Summary of Statistics for Selected Items

Distribution of item difficulty indexes. Requirements for the desired p-value distribution were met for all content areas except tool functions. The slight departure from the desired distribution observed in the tool functions items resulted from a shortage in the p-value range .46 - .55, and was the same for both forms (Table 2).

Internal consistency. The means of the internal consistency indexes for the items in each content area (Table 3) were essentially the same for the two forms ($\bar{r}_{bis}$ = .57 - .75 for AFQT 7 and .58 - .76 for AFQT 8). The serial order of magnitude of the indexes was the same in the two forms--verbal ability highest, tool functions lowest--as was the case also for earlier AFQT forms.

Relationship with operational AFQT score. The mean biserial coefficients for each content area were essentially the same on the two forms of AFQT (Table 3). Tool functions items showed the lowest correlation with operational AFQT score ($\bar{r}_{bis}$ = .37, .35).

Relationship among content areas. The two forms of AFQT showed essentially the same pattern of correlation among the content areas (Table 4). Verbal ability and arithmetic reasoning showed the closest relationship ($\bar{r}_{bis}$ = .58 - .62). Correlation involving tool functions items was lowest ($\bar{r}_{bis}$ = .28 - .44). Spatial relations showed a closer relationship to verbal ability and arithmetic reasoning ($\bar{r}_{bis}$ = .48 - .50) than to tool functions ($\bar{r}_{bis}$ = .34 - .44).

Table 2

DIFFICULTY INDEXES OF AFQT 7 AND 8 ITEMS
(p-values, corrected)

AFQT Percentile	Corrected p-value	Desired No. of items	Number of items selected									
			Verbal Ability		Arithmetic Reasoning		Tool Functions		Spatial Relations		Total Test	
			AFQT 7	AFQT 8	AFQT 7	AFQT 8	AFQT 7	AFQT 8	AFQT 7	AFQT 8	AFQT 7	AFQT 8
100 - 75	.00 - .25	1	1	1	1	1	1	1	1	1	4	4
74 - 65	.26 - .35	2	2	2	2	2	2	2	2	2	8	8
64 - 55	.36 - .45	2	2	2	2	2	3	3	2	2	9	9
54 - 45	.46 - .55	4	4	4	4	4	2	2	4	4	14	14
44 - 35	.56 - .65	4	4	4	4	4	5	5	4	4	17	17
34 - 25	.66 - .75	6	6	6	6	6	6	6	6	6	24	24
24 - 15	.76 - .85	4	4	4	4	4	4	4	4	4	16	16
14 - 5	.86 - .95	2	2	2	2	2	2	2	2	2	8	8
	Total	25	25	25	25	25	25	25	25	25	100	100
Mean corrected p-value			.62	.62	.61	,61	.62	.61	.61	.61	.61	.61

Table 3

MEAN INTERNAL CONSISTENCY INDEXES AND CORRELATION WITH OPERATIONAL AFQT

Content Area	AFQT 7		AFQT 8	
	Internal Consistency[a]	Operational AFQT[b]	Internal Consistency[a]	Operational AFQT[b]
Verbal Ability	.75	.61	.76	.62
Arithmetic Reasoning	.69	.57	.72	.59
Tool Functions	.57	.37	.58	.35
Spatial Relations	.68	.51	.66	.53

[a] $\bar{r}_{bis}$ between pass-fail on item and total score on all items of same content area.

[b] $\bar{r}_{bis}$ between pass-fail on item and total score on operational AFQT 5 and 6.

Table 4

MEAN BISERIAL COEFFICIENTS OF CORRELATION BETWEEN ITEMS IN ONE CONTENT AREA OF AFQT 7 AND 8 AND TOTAL SCORES ON EACH OF THE OTHER AREAS

	AFQT 7	AFQT 8
Verbal Ability Items vs		
Arithmetic Reasoning	.60	.62
Tool Functions	.39	.42
Spatial Relations	.49	.50
Arithmetic Reasoning Items vs		
Verbal Ability	.58	.59
Tool Functions	.37	.35
Spatial Relations	.49	.50
Tool Functions Items vs		
Verbal Ability	.32	.30
Arithmetic Reasoning	.30	.28
Spatial Relations	.35	.34
Spatial Relations Items vs		
Verbal Ability	.48	.48
Arithmetic Reasoning	.50	.50
Tool Functions	.42	.44

STANDARDIZATION AND EQUIVALENCE STUDIES

General Design and Sampling

The general design for the standardization of AFQT 7 and 8 was similar to the designs employed in the standardization of earlier forms of AFQT and in the development of item statistics for the experimental tests. The intent of the design was to convert the raw scores into national norms derived from the distribution of ability in the mobilization population. These norms are the Army standard scores (and their percentile equivalents) of the estimated distribution of AGCT scores of 11,694,229 enlisted men and officers in the Armed Forces as of December 1944, defined as the mobilization population and described by Uhlaner (1952). The basic technique has involved adjusting the scale of the new test to match the distribution of AGCT in samples in which scores on both tests are available.

In applying the technique, certain assumptions are necessary for the reason that complete population distributions for the two measures--the reference test and the experimental test--are not available. Even with large samples, it is quite possible that two measures of the same underlying ability may be skewed in opposite directions--another way of saying that one test may be easier than the other. Hence, if scores on the experimental test are to be meaningfully interchangeable with those on the reference test, it is necessary to assume that the distribution functions (mean, standard deviation, skewness, and kurtosis) for the converted scores on the two measures in the standardization sample are essentially identical. For this assumption, it is necessary to assume further that the sample represents the parent population equally well for both measures and that there is substantial correlation between the two measures in the parent population.

Again, all the services contributed examinees. The mobilization samples were established by the currently operational AFQT 5 and 6, and the selection of cases for analysis was without regard to testing location and service. To provide the basis for converting raw scores to mobilization percentiles, the AFQT Reference Test R 9 was administered. It was expected that by using one test for selection of cases and another for developing the norms, bias in the norms would be less than if one test were used for both purposes.

Data Collection

Tests were administered during June through August 1959[5] to a total of 3649 examinees, to provide 1000 complete cases for each form. To half the examinees the Reference Test R 9 was administered first; to the other half, one of the new forms of AFQT was administered first. Testing

[5] Data collection was protracted because of reduced input to AFES.

sessions were 2-1/2 hours long, with 40 minutes actual testing time for R 9, 50 minutes for AFQT. Tests for the equivalence study were administered at the same time and at the same locations (page 8) to an additional 1028 examinees to provide 600 complete cases. Both AFQT 7 and AFQT 8 were administered to all examinees, one-half having AFQT 7 administered first and the other half, AFQT 8. All tests were administered by regularly assigned enlisted examining personnel under the supervision of research personnel. Standard testing conditions were maintained. Quotas of examinees for each AFQT decile were established for each location.

At the training centers, the tests were administered during the early weeks of combat training to enlistees and inductees who had scored above the 30th percentile on AFQT (percentiles 31 - 100). Men were tested a few weeks after classification testing. Intervals as long as several months might have elapsed after operational AFQT testing at AFES. At AFES, only preinductees in the AFQT percentile interval 1 - 30 were examined. All failures included had been identified as true failures. The interval represented an expansion of the interval examined at AFES with the experimental tests. The change was occasioned by the supplementary screening of examinees obtaining percentile scores of 10 - 30 on AFQT, instituted by the Army in August 1958, a procedure which left only a biased portion of the 10 - 30 interval available for testing at training centers.

Selection of Cases for Standardization and Equivalence Studies

From the 1800 cases collected for each of the two forms in the standardization study, 1000 complete cases were selected for the analysis of data on each form in such a manner as to reproduce the mobilization population distribution. The selection was accomplished, as was the selection for the analysis of the experimental test data, by drawing proportional numbers of cases from each testing location and service. From the total sample, 100 cases were selected for each AFQT decile to produce a rectangular distribution of cases. The distribution within each decile was made up of 50 cases from each of the two orders of administration, rectangularly distributed.

In a similar manner, cases were selected for analysis of equivalence data. From the 1028 examinees tested, 300 cases were selected for each of the two orders of testing, each decile containing 30 rectangularly distributed cases for each order.

Table 5 identifies the various subsamples and summarizes the number of cases selected for the analysis.

Table 5

SELECTION OF CASES FOR AFQT 7 AND 8 STANDARDIZATION AND EQUIVALENCE ANALYSIS

Analysis	Subsample	N	Testing Order	
			Administered First	Administered Second
Standardization	1	500	R 9	AFQT 7
	2	500	AFQT 7	R 9
	3	500	R 9	AFQT 8
	4	500	AFQT 8	R 9
	5 (1 + 2)	1000	(AFQT 7 and R 9)	
	6 (3 + 4)	1000	(AFQT 8 and R 9)	
Equivalence	7	300	AFQT 7	AFQT 8
	8	300	AFQT 8	AFQT 7
	9 (7 + 8)	600	(AFQT 7 and 8)	

Development of Norms

Raw scores of AFQT 7 and 8 were converted to percentiles in the mobilization population. Frequency and cumulative frequency distributions of raw scores were prepared for AFQT 7 (subsamples 1 and 2, Table 5), for AFQT 8 (subsamples 3 and 4), and for AFQT 7 and AFQT 8 combined (subsamples 5 and 6). Corresponding distributions were prepared of the mobilization percentile equivalents to the raw scores on R 9. The R 9 percentiles are the equivalents of AGCT Army standard scores in the World War II mobilization population. Raw scores on both AFQT and R 9 were corrected for chance success: R - W/3. To each AFQT raw score was assigned the R 9 percentile score which had the same cumulative frequency in the sample as did the AFQT score. Since R 9 was administered to the same samples as were AFQT 7 and 8 (N = 1000 each), it was not necessary to compute percentages.

As usually occurs, the initial conversion of raw scores to scaled scores (mobilization population percentiles) did not yield complete correspondence between increments in the raw scores and increments in the scaled scores throughout the entire distribution. Particularly important were irregularities which provided no exact integral raw scores for particular percentile scores critical for operational use (percentiles 10, 31, 65, 93). Accordingly, adjustments were made to smooth the correspondence in the increments, particularly about the critical percentile scores, keeping as close as possible to the initial conversions provided by the data. Examination of the conversions for AFQT 7 and for AFQT 8 revealed their general similarity. Since the descriptive statistics also indicated the similarity of the two forms, the conversions based on the combined AFQT 7 and AFQT 8 distribution were prepared for operational use.

The major smoothing adjustments were systematic and involved the following explicit assumptions: (1) although scores are recorded as integral numbers, they reflect a continuous scale; (2) recorded scores are the midpoints of intervals; (3) in large samples, scores with zero frequencies occurring between scores with significant frequencies are sampling accidents or are artifacts of the scale or scoring system; and (4) with wide-range measures consisting of a substantial number of items, the curve of scale values against raw scores is smooth in a wide-range parent population; irregularities are artifacts resulting from the requirement for integral values in the final conversion table.

Descriptive Statistics

AFQT means, standard deviations, and coefficients of correlation with the reference tests were practically identical for all subsamples--that is, for different orders of administration (Table 6).

The four content areas were scored as subtests (R - W/3) and the product moment intercorrelations computed (Table 7). The highest correlation was between verbal ability and arithmetic reasoning. Correlation between tool functions and the other subtests was the lowest. Relationships shown are generally similar to those found in the analysis of the experimental tests as indicated by mean biserial correlation between scores on items in one content area and total scores on the 75 items in each of the other content areas (see Table 4). The intercorrelations were of sufficient magnitude to indicate a common relationship, yet not so great as to preclude unique contributions by each content area.

Correlation with Reference Test R 9 was of the same magnitude (r = .85 - .86) as in the standardization of previous forms of AFQT. This finding is of particular interest since R 9 was used here only to develop the norms and not, as previously, also to select cases for the standardization samples. Correlation between AFQT 7 and 8 and operational AFQT 5-6 was substantial (r = .89 - .90), a finding in line with other indications of satisfactory administration and scoring of operational AFQT.

Table 6

PRODUCT MOMENT CORRELATION OF AFQT 7 AND 8 WITH REFERENCE TEST R 9 AND OPERATIONAL AFQT

		AFQT 7 (raw score)		AFQT 8 (raw score)		R 9 (raw score)		AFQT 7 or 8 vs:	
Subsample	N	Mean	S.D.	Mean	S.D.	Mean	S.D.	R 9	Operational AFQT 5-6
1	500	61.2	23.2			53.4	25.2	.85	.89
2	500	60.7	23.4			52.2	23.4	.86	.90
3	500			59.2	23.2	53.4	24.5	.85	.89
4	500			60.3	23.0	53.2	24.4	.85	.90
5 (1 + 2)	1000	61.0	23.3			52.8	24.3	.87	.90
6 (3 + 4)	1000			59.8	23.1	53.3	24.5	.86	.89

Table 7

PRODUCT MOMENT INTERCORRELATIONS AMONG SCORES ON CONTENT AREAS--AFQT 7 AND 8

Content Area		Intercorrelation			
Verbal Ability	(V)	V			
Arithmetic Reasoning	(A)	.75	A		
Tool Functions	(T)	.44	.43	T	
Spatial Relations	(S)	.58	.63	.53	S

Item statistics were computed in the standardization samples and compared with item statistics for the same items available from the experimental tests. Several considerations impelled the comparison: Since the selected items represented a relatively small proportion of the total experimental items, internal consistency indexes and correlation with total operational AFQT could well be higher in the final tests (100 items) than in the experimental tests (300 items). On the other hand, since a number of items were rejected because they were surplus and not because they were considered less effective than the selected items, it was possible that the differences in item statistics would be minimal. Other considerations were also involved, including adjustments in the scores of the two lowest deciles to allow for terminal omissions, and the extent to which item correlation was affected by correlation of error compone ts.

Table 8 shows, for the experimental and standardization samples, means of p-values (corrected), internal consistency indexes, and correlation with the operational AFQT for each of the content areas. Taking the data at face value (no statistical tests of significance were applied), the final tests appear to be somewhat easier, more consistent internally, and better correlated with the operational AFQT than originally estimated. However, the net differences, assuming their significance, are slight.

No systematic examination was made of the differences for individual items. Some items, as might be expected, showed substantial differences between experimental and standardization values. In part, such differences might be expected to occur by chance, particularly as they might represent artifacts of uneven dichotomization. It is possible, nevertheless, that more systematic examination might be profitable with a view to improving item construction and selection methods.

Equivalence Results

The degree of equivalence of AFQT 7 and AFQT 8 was determined for the entire distribution and for particular critical scores. As already indicated (Table 6), the means, standard deviations, and correlation with the reference tests were practically identical for the two forms.

Equivalence was also measured in subsamples 7 and 8 to which both forms were administered, the subsamples differing only in the order of administration. No effects of test order were observed, as indicated by the similarity of means and standard deviations, as well as by the similarity of correlation between the two forms (r's = .94, .92). The correlation between the two forms computed for the combined sample (Sample 9 in Table 5) was r = .94, with standard errors of measurement (using .94 as the reliability coefficient) of 5.4 for AFQT 7 and 5.4 for AFQT 8 (Table 9). The substantial alternate form reliability and the similar standard errors of measurement indicate the essential equivalence of the two forms throughout the entire distribution of scores.

Table 8

SUMMARY OF STATISTICS FOR ITEMS IN FINAL FORMS OF AFQT 7 AND 8
BASED ON EXPERIMENTAL ITEM ANALYSIS AND
STANDARDIZATION SAMPLES

Content Area	Mean Corrected p-value				Mean Internal Consistency Index ($\bar{r}_{bis}$)				Mean Correlation with Operational AFQT ($\bar{r}_{bis}$)			
	AFQT 7		AFQT 8		AFQT 7		AFQT 8		AFQT 7		AFQT 8	
	Exp.	Final	Exp.	Final	Exp.	Final	Exp.	Final	Exp.	Final	Exp.	Final
Verbal	.62	.65	.62	.65	.75	.83	.76	.80	.61	.71	.62	.70
Arithmetic Reasoning	.61	.63	.61	.63	.69	.74	.72	.76	.57	.62	.59	.62
Tool Functions	.62	.63	.61	.64	.57	.62	.58	.62	.37	.40	.35	.37
Spatial Relations	.61	.64	.61	.60	.68	.71	.66	.69	.51	.54	.53	.51

Table 9

PRODUCT MOMENT COEFFICIENTS OF CORRELATION BETWEEN AFQT 7 AND AFQT 8
(Equivalence)

		AFQT 7 raw score		AFQT 8 raw score		AFQT 7 vs AFQT 8
Equivalence Subsamples	N	Mean	S.D.	Mean	S.D.	r
(AFQT 7 given first)	300	60.6	21.8	60.3	22.3	.94
(AFQT 8 given first)	300	60.7	22.1	59.4	22.2	.92
Combined sample	600	60.6	21.9	59.8	22.2	.94
SE (Measurement)		5.4		5.4		

To measure equivalence between the two forms at selected critical scores, tetrachoric correlation coefficients were computed, dichotomizing at percentiles 10, 21, 31, 51, 65, and 93. As might be expected from the high product moment correlation between AFQT 7 and AFQT 8 and the similar standard errors of measurement, the tetrachoric coefficients in the combined sample of 600 cases varied little (.94 - .97).

AFQT and Level of Education

The relationship between AFQT 7-8 and educational level was measurec by computing the product moment correlation between AFQT percentile scores and number of years of formal education (Table 10). The correlation (r = .53) was lower than that obtained for previous forms of AFQT and is in line with the requirement that performance on AFQT not be excessively dependent upon amount of formal education. Why this correlation should be lower than comparable correlations for the earlier forms of AFQT is not clear. One possible explanation is that in selecting items for AFQT 7 and 8, items highly correlated with amount of education happened to be rejected. Another is that with the passage of time, the level of education reached has become less related to level of ability. A third possible explanation is that this lower correlation is an artifact of sample composition. The samples were composed of both enlistees and inductees, who generally differ in educational level for comparable AFQT scores. If the proportions of enlistees and inductees differed from those in previous samples, or if the difference between enlistees and inductees in educational level was greater than in previous samples, the lower correlation between AFQT 7 and 8 and educational level could be expected.

Table 10

RELATIONSHIP BETWEEN AFQT 7-8 SCORES AND EDUCATIONAL LEVEL

AFQT 7-8 Percentiles	N	Years of Education Mean	S.D.
91 - 100	190	12.3	1.3
81 - 90	218	12.2	1.1
71 - 80	225	11.9	1.2
61 - 70	237	11.8	1.4
51 - 60	212	11.6	1.1
41 - 50	150	11.4	1.3
31 - 40	156	10.9	2.1
21 - 30	199	10.7	1.9
10 - 20	241	9.6	2.2
1 - 9	138	8.9	2.5
1 - 100	1966	11.2	1.9

Correlation between AFQT scores
and number of years of education: r = .53

Progressive increases in AFQT decile were accompanied by progressive increases in the mean number of years of education. Among AFQT failures (percentiles 1 - 9) were examinees who had as much schooling (11.4 years) as the average for examinees obtaining AFQT percentile scores 41 - 50. This point is of some operational significance since it indicates that some high school seniors and graduates may be found among legitimate AFQT failures. In the standardization samples, deliberate failures identified by terminal screening procedures were excluded (page 24); there is no reason to believe that many deliberate failures escaped detection.

Further, the lower deciles (1 - 40) varied more than did the upper deciles. This difference may reflect the virtual ceiling on educational level in preinduction samples imposed by deferment policies and in applicant-for-enlistment samples imposed by socio-economic conditions. However, it is possible that AFQT scores are underestimates of the ability of a number of those at the higher educational levels; or that sizable numbers are exposed to educational levels beyond their levels of ability. This second possibility is supported by evidence which indicates that AFQT and comparable classification tests are better predictors of military trainability than is number of years of formal schooling as reported by the examinees (Sharp, Helme, and White, April 1958; and Bayroff, Seeley, and Anderson, February 1960).

REFERENCES

Publications of the
U. S. Army Personnel Research Office, OCRD, DA

Bayroff, A. G. The mobilization base for AFQT norms. Research Memorandum 63-8. May 1963.

Bayroff, A. G., Morton, Mary A., Anderson, A. A., and Hilligoss, R. E. Item analysis and selection of items for standardization forms of AFQT 7 and 8. Research Memorandum 60-10. April 1960.

Bayroff, A. G., Seeley, L. C., and Anderson, A. A. Relationship of AFQT to rated training performance. Technical Research Note 106. February 1960.

Bayroff, A. G., Thomas, J. A., Kehr, Carol J. Evaluation of EST for predicting AFQT performance. Technical Research Report 1114. February 1959.

Morton, Mary A., Houston, T. C., and Bayroff, A. G. Development of Enlistment Screening Test, Forms 3 and 4. Technical Research Report 1102. May 1957.

Sharp, L. H., Helme, W. H. and White, R. K. Prediction of success in selected electronics repair jobs. Technical Research Note 92. April 1958.

Uhlaner, J. E. Development of Armed Forces Qualification Test and predecessor Army Screening Test, 1946-1950 (3d printing). Technical Research Report 976. November 1952.

Additional References

Anderson, A. A. Sample bias by eliminating incomplete answer sheets. *American Psychologist*, *15*, 446, 1960.

Terminal Screening Guide. Department of the Army Pamphlet 611-37. July 1956.

AD 23/1, 28/4

U. S. Army Personnel Research Office, OCRD, DA
DEVELOPMENT OF ARMED FORCES QUALIFICATION TEST 7 AND 8 by
A. G. Bayroff and Alan A. Anderson. May 1963. Rept. on Input
Quality 00-01 Proj.--39 p. incl tables, figures 28 Ref.
(USAPRO Technical Research Report No. 1132)
(DA Project 2J024701A713) Unclassified Report

The Armed Forces Qualification Test, the screening test used by all the services, must provide both a measure of general military trainability and measures of specific aptitudes. Following the research design for previous forms, experimental test items in four content areas developed by the separate services were administered to 3000 Armed Forces personnel for item analysis and item selection. Final forms were then administered to standardization samples representative of the mobilization population as a basis for conversion of test scores to percentile norms. AFQT 7 and 8 correlated substantially with preceding operational forms (r = .89 - .90) and are satisfactory alternate forms for screening. Correlation of AFQT 7-8 with years of formal education (r = .53) was slightly less than for the previous forms. Because of the high degree of equivalence of the two forms (r = .94) established in samples totaling 600 cases, a single conversion table was established for AFQT 7 and 8. Based on experimentation, instructions for administering AFQT 7 and 8 have been made shorter and simpler than for previous forms, with no loss in test effectiveness.

UNCLASSIFIED
Human Resources Research
--Personnel Classification

AD 23/1, 28/4

U. S. Army Personnel Research Office, OCRD, DA
DEVELOPMENT OF ARMED FORCES QUALIFICATION TEST 7 AND 8 by
A. G. Bayroff and Alan A. Anderson. May 1963. Rept. on Input
Quality 00-01 Proj.--39 p. incl tables, figures 28 Ref.
(USAPRO Technical Research Report No. 1132)
(DA Project 2J024701A713) Unclassified Report

The Armed Forces Qualification Test, the screening test used by all the services, must provide both a measure of general military trainability and measures of specific aptitudes. Following the research design for previous forms, experimental test items in four content areas developed by the separate services were administered to 3000 Armed Forces personnel for item analysis and item selection. Final forms were then administered to standardization samples representative of the mobilization population as a basis for conversion of test scores to percentile norms. AFQT 7 and 8 correlated substantially with preceding operational forms (r = .89 - .90) and are satisfactory alternate forms for screening. Correlation of AFQT 7-8 with years of formal education (r = .53) was slightly less than for the previous forms. Because of the high degree of equivalence of the two forms (r = .94) established in samples totaling 600 cases, a single conversion table was established for AFQT 7 and 8. Based on experimentation, instructions for administering AFQT 7 and 8 have been made shorter and simpler than for previous forms, with no loss in test effectiveness.

UNCLASSIFIED
Human Resources Research
--Personnel Classification

AD 23/1, 28/4

U. S. Army Personnel Research Office, OCRD, DA
DEVELOPMENT OF ARMED FORCES QUALIFICATION TEST 7 AND 8 by
A. G. Bayroff and Alan A. Anderson. May 1963. Rept. on Input
Quality 00-01 Proj.--39 p. incl tables, figures 28 Ref.
(USAPRO Technical Research Report No. 1132)
(DA Project 2J024701A713) Unclassified Report

The Armed Forces Qualification Test, the screening test used by all the services, must provide both a measure of general military trainability and measures of specific aptitudes. Following the research design for previous forms, experimental test items in four content areas developed by the separate services were administered to 3000 Armed Forces personnel for item analysis and item selection. Final forms were then administered to standardization samples representative of the mobilization population as a basis for conversion of test scores to percentile norms. AFQT 7 and 8 correlated substantially with preceding operational forms (r = .89 - .90) and are satisfactory alternate forms for screening. Correlation of AFQT 7-8 with years of formal education (r = .53) was slightly less than for the previous forms. Because of the high degree of equivalence of the two forms (r = .94) established in samples totaling 600 cases, a single conversion table was established for AFQT 7 and 8. Based on experimentation, instructions for administering AFQT 7 and 8 have been made shorter and simpler than for previous forms, with no loss in test effectiveness.

UNCLASSIFIED
Human Resources Research
--Personnel Classification

AD 23/1, 28/4

U. S. Army Personnel Research Office, OCRD, DA
DEVELOPMENT OF ARMED FORCES QUALIFICATION TEST 7 AND 8 by
A. G. Bayroff and Alan A. Anderson. May 1963. Rept. on Input
Quality 00-01 Proj.--39 p. incl tables, figures 28 Ref.
(USAPRO Technical Research Report No. 1132)
(DA Project 2J024701A713) Unclassified Report

The Armed Forces Qualification Test, the screening test used by all the services, must provide both a measure of general military trainability and measures of specific aptitudes. Following the research design for previous forms, experimental test items in four content areas developed by the separate services were administered to 3000 Armed Forces personnel for item analysis and item selection. Final forms were then administered to standardization samples representative of the mobilization population as a basis for conversion of test scores to percentile norms. AFQT 7 and 8 correlated substantially with preceding operational forms (r = .89 - .90) and are satisfactory alternate forms for screening. Correlation of AFQT 7-8 with years of formal education (r = .53) was slightly less than for the previous forms. Because of the high degree of equivalence of the two forms (r = .94) established in samples totaling 600 cases, a single conversion table was established for AFQT 7 and 8. Based on experimentation, instructions for administering AFQT 7 and 8 have been made shorter and simpler than for previous forms, with no loss in test effectiveness.

UNCLASSIFIED
Human Resources Research
--Personnel Classification